Vowel and Speaker Identification in Natural and Synthetic Speech*

Ilse Lehiste and David Meltzer**

Department of Linguistics and Department of Electrical Engineering

The purpose of this study was to develop a simple means for evaluating the relative quality of synthesizers. Several considerations had to be kept in mind: the synthesis should involve standardized materials and yet be adaptable to each new situation; testing procedures should be easy to follow and to replicate; and the results of the evaluation of different synthesizers should be comparable. We decided to use a set of vowels, since vowels are usually simpler to synthesize, and the results are more likely to reflect the performance of the synthesizer than the skill of the person performing the synthesis.

Method.

We selected a set of ten monophthongal American English vowels: /i ɪ ɛ æ ə ɔ U u ʌ ɝ/. The vowels were produced by a male speaker, a female speaker, and a child. The two adult speakers were phoneticians; the child repeated the productions of its mother (a phonetician). The vowels were recorded on magnetic tape in an anechoic chamber, analyzed spectrographically, and randomized for a listening test. The first listening tape contained one production of each of the vowels in random order, for a total of 30 items.

All 30 vowels were synthesized on a Glace-Holmes synthesizer,[1] using formant positions measured from spectrograms obtained from the first set. Fundamental frequency values were taken from averages for these vowels published by Peterson and Barney.[2] A second listening tape was prepared, on which the synthesized 30 vowels appeared in random order, with 5-second intervals.

A third set of vowels was generated on the basis of average formant and fundamental frequency values published by Peterson and Barney. In this set, formant values for men, women and children were combined with the respective fundamental frequencies, resulting in 9 different combinations for each of the ten vowels. Three of the nine sets are directly comparable to the materials contained on the first two tapes; six represent combinations which do not occur in normal speech. These combinations were synthesized to gain some information about the relative importance of formant structure and fundamental frequency in the identification

of speakers and vowels. The third set of vowels, 90 items in all, was randomized and re-recorded in the same manner as the second set.

Six listening tapes were prepared, containing all 150 stimuli. On each tape, the order of the three sets of vowels was varied, so that the effects of order of presentation of normal vowels and different synthetic vowels would be equalized. The tapes were presented to 10 listeners each, for a total of 60 listeners. The listeners had had approximately three months' training in (English) phonetics and were familiar with the phonetic symbols. The task of each listener was to identify both the vowel and the speaker by placing the proper phonetic symbol in one of three columns, thus assigning the vowel to a male speaker, female speaker, or a child.

Results.

The results of the listening tests are presented in Tables 1-7. Table 1 presents the results of vowel identification for normal productions. Table 2 gives comparable data for the set of vowels synthesized on the basis of measurements made from the first set.

Table 3 presents comparable data for the set of vowels synthesized on the basis of the Peterson-Barney averages. This table contains only normal combinations, i.e. male formants and fundamental frequency, female formants and fundamental frequency, and child's formants and fundamental frequency. It is thus directly comparable to Tables 1 and 2. Table 4 summarizes the vowel and speaker identification data for these three sets of vowels: normal productions, synthesis from measured values (attempting to recreate the first set synthetically), and synthesis from average values.

Tables 5, 6 and 7 present data for the set of 90 vowels synthesized on the basis of averages. The tables contain information obtained for all nine possible combinations of formant and fundamental frequencies. Table 5 presents speaker identification scores. Table 6 was generated by averaging Table 5 results across fundamental frequency changes and across formant structure changes. Table 7 gives vowel identification scores.

Discussion.

1. Speaker Identification.

A first observation is that male speakers are identified more easily than women and children, who are frequently confused with each other. This would seem to be a trivial observation; it is interesting, however, that the confusions are much greater in synthesized sets than in the normal set. Evidently the normal productions contain some additional information which is used by listeners in making the decision, and which is not reproduced on

the Glace-Holmes synthesizer.

Tables 5 and 6 show that formant structure is a relatively more important cue in speaker identification than fundamental frequency. For example, vowels produced with male formants, but female fundamental frequency, were assigned to a male speaker in 80.8% of instances, while vowels synthesized with female formants, but with male fundamental frequency, were assigned to a male speaker in only 18.6% of the cases.

2. Vowel Identification.

First of all, it is obvious that children's vowels are relatively difficult to identify. In the case of the first two sets (Tables 1 and 2), one might attribute this to the fact that the child whose recording of the vowels was used in this test may not have succeeded in pronouncing the vowels correctly. But a comparison with synthesis from the Peterson-Barney averages (Table 3) shows that this is not so: here, too, the score for children's vowels was the lowest, and the reason must be sought elsewhere. A simple answer might be provided by observing that children's formants are usually not well defined, since the high fundamental frequency of a child's voice would furnish only one or two harmonics per formant. If this is the true reason, the identifiability of a child's vowels should increase when a man's fundamental frequency is used. Table 7 shows that this is not the case: children's formants, with a male fundamental frequency, resulted in an average vowel identification score of 43.9%, compared to 67.9% for children's formants combined with children's fundamental frequency.

It is noticeable also that synthesis from averages produced relatively higher vowel identification scores than synthesis from measurements of the normal set. A possible reason is that Peterson and Barney used for their averages only vowels that had been correctly identified by a panel of listeners, and discarded those that were not unanimously accepted. Thus the Peterson-Barney averages represent some kind of idealized vowels--not what an average speaker would produce, but what an average listener would accept.

Vowels obviously differ a great deal in their relative identifiability. In normal productions, the vowels /U/ and /ʌ/ had the lowest scores. Surprisingly, /ʌ/ had a relatively high score in the synthetic set based on measurements; in this set, the lowest scores were obtained for /ɛ/ and /U/. For the set of vowels synthesized from averages, the lowest scores were associated with /ʌ/ and /U/, as had been the case with the normal set. High front vowels and /ɝ/ had consistently high identification scores.

A surprising result was the low identification score of /i/ in the set synthesized from measurements (Table 2). We hypothesize that this might be due to the fact that the fourth

formant was not used in the synthesis; however, /i/ synthesized on the basis of averages received a high score, even though F4 was not used either. The relatively high score for the child may be explained by the fact that a modification was introduced into the synthesizer to obtain the characteristic high third formant for the child's /i/, which would otherwise have been out of range of the Glace-Holmes synthesizer.

An analysis of the substitutions made by the listeners would add some interesting information, but would contribute little to the primary aim of the study: establishing an evaluation measure for synthesizers.

We propose to use the difference between normal scores and scores obtained with synthetic vowels as an evaluation measure. The use of the Peterson-Barney data will provide a fixed reference. For the current state of our Glace-Holmes synthesizer, we have to evaluate its performance as approximately 25% below normal speech. This is based on a comparison of overall scores. The overall vowel identification score for the normal set (all three speakers combined) was 79.46%; the overall speaker identification score (all ten vowels combined) was 90.03%. The corresponding scores for the set synthesized from measured spectrograms were 50.87% and 69.73% respectively. The differences between the scores obtained for the normal set and the synthesized set were -28.59% for vowel identification and -20.30% for speaker identification, giving an approximate degradation of the signal of 25%. Compared with the synthesis from averages, the performance of the Glace-Holmes synthesizer is much better: the difference for vowel identifications between the normal set and the synthesis from averages was -4.23%, and for speaker identification, -15.76%, for an average degradation of 10%.

Footnotes

*This research was supported in part by the National Science Foundation under Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University. The paper was presented at the 82nd meeting of the Acoustical Society of America, Denver, Colorado, October 21, 1971. The authors are indebted to Dr. G. Powers of the Speech Department of The Ohio State University for his help in carrying out the listening tests.

**David Meltzer is currently with the I.B.M. Corporation, Poughkeepsie, New York.

[1]Glace, Donald A. A Parallel Resonance Synthesizer for Speech Research. Unpublished Manuscript.

[2]Peterson, Gordon E., and Harold L. Barney (1952) "Control Methods Used in a Study of the Vowels." JASA 24.175-184.

TABLE 1
VOWEL IDENTIFICATION: PHONATED VOWELS, NORMAL SPEAKERS
Scores given in per cent correct

Vowel	Male	Female	Child	Overall correct (vowel & speaker)	Overall correct (MFC combined)
i	100	90	72	87.33	93.00
ɪ	96	74	87	85.67	93.67
ɛ	70	81	97	82.67	85.67
æ	96	77	90	87.67	73.67
a	94	57	25	58.67	73.67
ɔ	81	67	64	70.67	77.33
U	80	63	10	51.00	56.33
u	98	75	90	87.67	98.67
ʌ	72	54	0	42.00	48.33
ɝʌ	96	78	31	68.33	74.33
Average	88.3	71.6	56.6	72.17	79.46

TABLE 2
VOWEL IDENTIFICATION: SYNTHESIZED VOWELS, BASED ON MEASUREMENTS OF PRODUCTIONS OF NORMAL SPEAKERS
Scores given in per cent correct

Vowel	Male	Female	Child	Overall correct (Vowel & speaker)	Overall correct (MFC combined)
i	8	8	54	23.33	36.00
ɪ	12	21	14	15.67	19.67
ɛ	19	52	12	27.67	39.67
æ	79	65	70	71.33	93.00
ɑ	46	45	42	44.33	68.00
ɔ	47	42	4	31.00	44.33
ʊ	10	4	24	12.67	19.33
u	49	12	8	23.00	34.33
ʌ	73	30	44	49.00	72.33
ɝ	50	74	38	54.00	82.00
Average	39.3	35.3	31.0	35.2	50.87

TABLE 3
VOWEL IDENTIFICATION: SYNTHESIZED VOWELS, FORMANT STRUCTURE AND FUNDAMENTAL FREQUENCY BASED ON AVERAGES GIVEN BY PETERSON & BARNEY (1952)
Scores given in per cent correct

Vowel	Male	Female	Child	Overall correct (Vowel & speaker)	Overall correct (MFC combined)
i	84	62	62	69.33	88.67
ɪ	76	47	70	64.33	77.00
ɛ	81	53	68	67.33	83.00
æ	79	60	72	70.33	89.67
a	60	56	53	56.33	76.67
ɔ	67	42	20	43.00	59.67
U	67	29	20	38.67	58.00
u	86	49	21	52.00	76.33
ʌ	37	38	29	34.67	47.00
ɝ	98	65	57	73.33	96.33
Average	73.5	50.1	47.2	56.93	75.23

TABLE 4
OVERALL SPEAKER AND VOWEL IDENTIFICATION
Scores given in per cent correct

Stimulus type	Speaker identification			Overall speaker identification score	Overall vowel identification score
	Male	Female	Child		
Normal speakers					
Male	99.2	0.4	0.4		
Female	2.2	81.0	16.8		
Child	0.0	10.1	89.9		
				90.03	79.46
Synthesis from measurements					
Male	96.2	3.0	0.8		
Female	9.8	62.2	28.0		
Child	5.2	44.0	50.8		
				69.73	50.87
Synthesis from averages					
Male	94.0	2.7	3.3		
Female	9.4	60.6	30.0		
Child	4.7	27.1	68.2		
				74.27	75.23

TABLE 5
SPEAKER IDENTIFICATION: SYNTHESIZED VOWELS, FORMANT STRUCTURE AND FUNDAMENTAL FREQUENCY BASED ON AVERAGES GIVEN BY PETERSON AND BARNEY (1952)
All vowels combined. Scores given in per cent correct.

Formants	Fundamental frequency	Identified as		
		Male	Female	Child
Male	Male	94.0	2.7	3.3
	Female	80.8	10.4	8.8
	Child	69.7	11.4	18.9
Female	Male	18.6	50.5	30.9
	Female	9.4	60.6	30.0
	Child	7.2	43.2	49.6
Child	Male	11.2	39.6	49.2
	Female	7.5	44.3	48.2
	Child	4.7	27.1	68.2
Average		33.68	32.20	34.12

TABLE 6
SPEAKER IDENTIFICATION, BASED ON A) FUNDAMENTAL FREQUENCY AND B) FORMANT STRUCTURE
All vowels combined. Scores given in per cent correct

		Identified as		
		Male	Female	Child
Fundamental frequency	Male	41.27	30.93	27.80
	Female	32.57	38.43	29.00
	Child	27.20	27.23	45.57
Formants	Male	81.50	8.17	10.33
	Female	11.73	51.43	36.84
	Child	7.80	37.00	55.20

TABLE 7

VOWEL IDENTIFICATION: SYNTHESIZED VOWELS, FORMANT STRUCTURE AND FUNDAMENTAL FREQUENCY BASED ON AVERAGES GIVEN BY PETERSON AND BARNEY (1952)

Scores given in per cent correct.

Formants	Fundamental frequency	i	ɪ	ɛ	æ	a	ɔ	ʊ	u	ʌ	ɝ	Average
Male	Male	88	78	81	86	64	67	69	86	41	100	76.0
	Female	96	66	73	96	87	78	61	83	46	80	76.6
	Child	78	31	40	89	83	44	10	14	26	19	43.4
Female	Male	88	54	54	53	52	22	37	40	42	96	53.8
	Female	98	76	91	91	84	73	66	83	58	98	81.8
	Child	78	31	40	89	83	44	10	14	26	19	43.4
Child	Male	83	35	24	3	25	17	46	75	35	96	43.9
	Female	98	70	87	92	47	61	63	72	85	96	77.1
	Child	80	77	77	92	82	39	39	60	42	91	67.9
Average		87.44	57.56	63.0	76.78	67.44	49.44	44.56	58.56	44.56	77.22	62.66